CHAPTER TWO

DEVELOPMENT AND ASSESSMENT OF MICROARRAY-BASED DNA FINGERPRINTING IN *EUCALYPTUS GRANDIS*

Published: *Theoretical and Applied Genetics* (2004) 109: 1329-1336

ABSTRACT

Development of improved *Eucalyptus* genotypes involves the routine identification of breeding stock and superior clones. Currently, microsatellites and random amplified polymorphic DNA (RAPD) markers are the most widely used DNA-based techniques for fingerprinting of these trees. While these techniques have provided rapid and powerful fingerprinting assays, they are constrained by their reliance on gel or capillary electrophoresis, and therefore, relatively low throughput of fragment analysis. In contrast, recently developed microarray technology holds the promise of parallel analysis of thousands of markers in plant genomes. The aim of this study was to develop a DNA fingerprinting chip for *Eucalyptus grandis*, and to investigate its usefulness for fingerprinting of eucalypt trees. A DarT-prototype chip was prepared using a partial genomic library from total genomic DNA of 24 *Eucalyptus grandis* trees, of which 22 were full siblings. A total of 384 cloned genomic fragments were individually amplified and arrayed onto glass slides. DNA fingerprints were obtained for 17 individuals by hybridising labeled genome representations of the individual trees to the 384-element chip. Polymorphic DNA fragments were identified by evaluating the binary distribution of their background-corrected signal intensities across full-sib individuals. Among 384 DNA fragments on the chip, 104 (27%) were found to be polymorphic. Hybridisation of these polymorphic fragments was highly repeatable ($R^2 > 0.91$) within the *E. grandis* individuals and they allowed us to identify all 17 full-sib individuals. Our results suggest that DNA microarrays can be used to effectively fingerprint large numbers of closely related *Eucalyptus* trees.

INTRODUCTION

Eucalyptus spp. are widely planted as exotics in many tropical and subtropical regions of the world (Eldridge et al. 1993). Since many of these plantations are commonly developed using vegetative propagation, the routine identification of clones and selection of elite genotypes has become increasingly important. Until recently, tree breeders have had to rely on detailed pedigree information and careful labeling to identify individual trees in breeding programs. However, incorrect identification is common and poses a major problem in forestry operations (Keil et al. 1994). DNA-based molecular markers have provided a solution to this problem. Several studies have thus shown that individual genotypes can be discriminated using molecular markers (Epplen et al. 1991; Nybom1991; Weising et al. 1991).

A variety of molecular marker techniques can be used for DNA fingerprinting. These techniques include restriction fragment length polymorphisms (RFLPs, Botstein et al. 1980), simple sequence repeats (SSRs, Weber and May 1989), random amplified polymorphic DNAs (RAPDs, Williams et al.1990), and amplified fragment length polymorphisms (AFLPs, Vos et al. 1995). Despite the high throughput afforded by some of these methods, they are all constrained by their dependence on gel electrophoresis. This hampers the processing of a large number of samples or markers in parallel (Smith and Beavis 1996). Furthermore, several of these methods require processing with many independent restriction enzymes or probes to achieve low error rates.

Originally designed for analysis of gene expression, DNA microarrays permit the parallel processing of large numbers of DNA fragments immobilised on a solid-state surface (Schena et al. 1995). To adopt microarray technology for fingerprinting and diversity studies, Jaccoud et al. 2001 recently reported the development of Diversity Array Technology (DArT™) in rice, while Borevitz et al. 2003 reported the use of oligonucleotide arrays to detect and genotype single feature polymorphisms (SFP) in *Arabidopsis*. No oligonucleotide arrays are available for *Eucalyptus* and therefore, the Diversity Array technique is the only microarray-based genotyping method that would be applicable for these trees. DArT™ is a solid state fingerprinting technique based on AFLP and enables analysis of large numbers of marker loci without any DNA sequence information. Microarray-based genotyping as implemented in the DArT™ technique is a 2-dye approach and relies on the detection of DNA fragments in a complex mixture of selectively amplified restriction fragments. Reduction of complexity by selective amplification allows comparison of polymorphic fragments among genotypes. This is achieved by hybridising DNA to an array containing a large number of DNA fragments, derived from genomic representations of an organism. However, plant genomes contain large amounts of highly repetitive DNA sequences and it is not clear how this feature might affect the rigor of hybridisation-based fingerprinting.

The aim of this study was to develop a prototype microarray chip to evaluate the potential of DNA microarrays for fingerprinting closely related *Eucalyptus* clones. In this study, the reproducibility of microarray hybridisation

profiles in *Eucalyptus grandis* was evaluated and recommendations for using this technology in plantation forestry were provided.

MATERIALS AND METHODS

Plant material and DNA extraction

A total of 15 full-sib progeny of *E. grandis* clone ZG14 (Mondi Forests, South Africa) were fingerprinted in this study. Clone ZG14 was used in a controlled cross with *E. grandis* clone TAG-S (Mondi Forests), from which 22 cloned progeny (clones 44D, 32A, 67D, 36E, 31C, 62D, 74C, 53B, 12C, 17C, 10D, 28D, 18C, 60D, 30B, 4D, 13C, 44C, 17D, 74C, 16C and 56E) were selected for the generation of a genomic representation of the whole full-sib family (described below). Genomic DNA was extracted from tree ZG14 and one ramet of each tree as described by Murray and Thompson (1980). The second parent tree (TAG-S) was lost during the early stages of this study and plant material was not available for it. A DNA sample was, therefore, obtained from tree TAG-5, a putative sibling relative of TAG-S.

Generation of genome representations

The method used for preparation of genome representations (Figure 1) was essentially the same as that described by Jaccoud et al. (2001). DNA samples were

pooled from 23 trees (144 ng DNA in total from 22 full-sib progeny and parental tree ZG14). The DNA in the pool was digested with 20 U *Pst*I using buffer H (Roche Diagnostics GmbH, Mannheim, Germany) in a reaction volume of 50 µL. The reactions were incubated at 37°C for 3 h and the restriction enzyme removed using an equal volume of phenol:chloroform. The DNA fragments were then precipitated with 100% EtOH and 100 mM NaCl. The precipitated DNA was washed with 70% ETOH and resuspended in 20 µl deionized water to a final concentration of 30 ng/µl.

Purified DNA was ligated to *Pst*I-specific adapters (Jaccoud et al. 2001) in a total volume of 30 µl at 10°C, overnight. The ligation mixture consisted of 1X ligation buffer, 2 U T_4 DNA ligase (Roche Diagnostics GmbH), 10 ng/µL l BSA (Amersham Biosciences, Piscataway, USA), 1.0 mM ATP (Amersham Biosciences) and 10 µM *Pst*I adapters. After ligation, 0.2 mM EDTA was added and the samples were heated at 70°C for 5 min to inactivate the ligase. The mixture was then diluted to 100 µl with water and 2 µl used as a template in a subsequent selective PCR reaction.

The PCR was performed in 50 µl containing 0.8 µM PCR primer (adapter +T), 0.25 mM of each dNTP, 1 U *Taq* polymerase, and 1 x reaction buffer (Roche Diagnostics GmbH). The PCR amplification consisted of 30 cycles of 94°C for 30 sec, 53°C for 45 sec, and 72°C for 1 min, with an initial denaturation step of 94°C for 5 min, and a final extension step of 72°C for 8 min.

Cloning, PCR amplification and sequencing of genomic fragments from representations

The amplified products were inserted into the PCR 2.1-TOPO vector using a T/A cloning kit (Invitrogen, Carlsbad, California, USA). After transforming *Escherichia coli* TOP 10F' host cells with ligation products, single colonies were grown overnight at 37°C in LB medium containing 50 µg/ml ampicillin. Recombinant *E. coli* clones were diluted in 1 vol of 50% glycerol and stored at -80°C. From each culture, 10 µl were transferred to 10 µl water and boiled for 10 min to disrupt the cells and release plasmid DNA into the growth medium. A 1 µl aliquot of this solution was used in a 100 µL PCR reaction with M13 forward (-20) and M13 reverse primers (Invitrogen). The reaction mix contained 1X PCR buffer, 1 U *Taq* polymerase (Roche Diagnostics GmbH), 0.25 mM of each dNTP, and 0.4 µM of each primer. The PCR amplification consisted of 30 cycles of 94°C for 30 sec, 53°C for 30 sec, and 72°C for 1 min, with an initial denaturation step of 95°C for 5 min, and a final extension step of 72°C for 7 min. Aliquots of the PCR products were separated on a 1.4% agarose gel for quality control. The remainder of each sample was then precipitated in 90% ethanol and 0.9 mM NaAc (pH 5.2) to exclude low molecular weight fragments. The precipitate was collected by centrifugation at 3600 x *g* for 30 min. Pellets were washed in 70% ethanol, dried, and then resuspended in deionized water at ~ 250 ng/µl.

Out of 384 amplified clones, forty random clones were sequenced. The insert sequences were subjected to similarity searches in GenBank using BLASTN and BLASTX. BLAST alignments were used to estimate the number of repetitive

clones in the library that could result in cross-hybridisation or uninformative spots on the array.

Array printing and processing

Equal volumes (10 µL each) of purified PCR product and 100% DMSO were transferred into a 384-well plate (Amersham Pharmacia Biotech). Eight replicates per fragment were arrayed on each slide at 250 µm spacing onto Vapour Phase Coated Glass Slides (Amersham Pharmacia Biotech) using a Molecular Dynamics Gen III spotter at the African Centre for Gene Technologies (ACGT) Microarray Facility, University of Pretoria, Pretoria, South Africa (http://fabinet.up.ac.za/microarray). Following printing, the slides were allowed to dry at 45-50% relative humidity overnight. Spotted DNA was then bound to the slides by UV-crosslinking at 250 mJ and baking at 80°C for 2 h.

Preparation of labeled probes

For microarray hybridisations, genome representations from parent tree ZG14 and 15 full-sib progeny were used. Tree TAG-5, the putative relative of parent TAG-S, was also included. Probe DNA from individual plants was prepared by restriction enzyme digestion of genomic DNA (144 ng per tree), ligation of restriction fragments to adapters, and subsequent amplification following the protocol described above. Amplicons were precipitated in one volume isopropanol to remove excess dNTPs. Labeling of the amplified fragments was carried out using the Klenow fragment of DNA Polymerase I (Roche Diagnostics GmbH). Each

labeling reaction contained 5 µg amplified DNA, 1.8 mM dNTP mix (0.3 mM of dATP, dGTP, dCTP each, 0.8 mM of dTTP, 0.1 mM Cy3-dUTP (Amersham Biosciences, Buckinghamshire, UK), 1 x hexanucleotide mix (Roche Diagnostics, GmbH) and 8 U Klenow enzyme (Roche Diagnostics GmbH). The reaction was incubated at 37°C overnight. After labeling, the DNA was column-purified (QIAquick PCR purification Kit, Qiagen GmbH, Germany).

Hybridisation and washing

Microarray slides were pre-hybridised for 20 min at 60°C in a solution containing 3.5 x SSC, 0.2 % SDS and 1% BSA (Roche Diagnostics GmbH). Slides were rinsed three times in deionized water and dried with N_2 gas. The Cy3-labeled probe was then dissolved in hybridisation solution containing 50% formamide (SIGMA), 25% 2 x hybridisation buffer (Amersham Pharmacia Biotech), and 25% deionized water. The mixture was denatured at 92°C for 5 min and quickly cooled on ice. The denatured probe (approximately 35 µl) was pipetted directly onto the microarray surface and covered with a glass coverslip (24 mm x 60 mm, No.1, Marienfeld, Germany). Slides were placed in a custom made hybridisation chamber (N. B. Engineering Works, Pretoria, South Africa) and incubated for 16 -18h in a 42°C water bath.

After hybridisation, slides were washed once in 1 x SSC, 0.2% SDS at 37°C for 4 min, twice in 0.1 x SSC, 0.2% SDS at 37°C for 4 min, twice in 0.1 x SSC at room temperature for 1 min, and then rinsed in deionised water for 2 seconds. Slides were dried using N_2-gas.

Scanning, image processing and data analysis

Slides were scanned using a GenePix 4000B Scanner (Molecular Dynamics, USA). The mean pixel intensity within each spot and the local background the spot were determined using Array Vision 6.0 software (Imaging Research Inc., Molecular Dynamics, USA). All signal intensities were background corrected. Abnormal spots (e.g. high background, dust, irregularities, etc.) were manually flagged for removal. Anomalous spots not detected through manual inspection were removed if the signal intensity of such spot was greater than 10% of the mean of the eight replicates on each slide. The mean background-corrected spot intensity of the remaining replicates of each DNA fragment was used in subsequent data analyses.

The single dye (Cy3) data were normalised across slides by regression on the spot intensity data for tree ZG14, which was used as a reference for normalisation of all progeny data. The normalised data were then converted into $\log_2$ intensity values.

Identification of polymorphic fragments

Polymorphic DNA fragments were identified in Microsoft Excel based on the bimodal distribution of their normalised intensity values across slides, consistent with their segregation as dominant PCR-based testcross (Aa:aa = 1:1) or intercross (A:aa = 3:1) markers. Relative intensity values were obtained by scaling the signal intensities to that of the DNA fragment with the highest intensity value across

slides (set to 1.0). The ranked spot intensities were plotted for each DNA fragment, and identification of DNA fragments with bimodal distribution was based on the presence of two clearly defined intensity classes with mean relative intensity values differing by at least 0.5. A binary scoring table of polymorphic spots was developed for all the *Eucalyptus* trees analyzed. The data for all the polymorphic spots were used to calculate the relative "distances" between the hybridisation profiles of individual *Eucalyptus* trees using Spearman correlation and hierarchical clustering (CLUSTER, available at http://rana.lbl.gov/). The clustering results were visualized with TREEVIEW (Eisen et al. 1998).

Verification of DNA polymorphisms

Two of the DNA polymorphisms detected in the array experiment were analysed by Southern hybridisation. *Pst*I digested total genomic DNA of nine individual trees were resolved on agarose gels and transferred to nylon membranes. Probes representing two of the polymorphic DNA fragments were labeled and hybridised to the *Pst*I digested DNA on the nylon membranes using the DIG High Prime DNA Labeling and Detection Starter Kit I (Roche Diagnostics GmbH, Germany).

Reproducibility of DNA microarray fingerprints

Tests were done on the reproducibility of hybridisation profiles starting from independently prepared genome representations, and that of stripping and re-hybridisation of the same slides. Repeated stripping and re-hybridisation of slides

allows for multiple rounds of hybridisation on the same slides. To test the reproducibility of the hybridisation fingerprints obtained from stripped slides, slides were treated using the protocol of Dolan et al. (2001) with minor modifications. Used slides were immersed four times in stripping buffer (2.5 mM Na_2HPO_4, 0.1% SDS) at 95°C for 25 s. Slides were then washed in deionised water at room temperature for 2 seconds and dried using N_2-gas. Stripped slides were scanned to verify that all signal had been removed. The stripped slides were then used for a repeat of the same hybridisation as before, but with independently labeled DNA. Data analysis was performed as described above. Independent replicates were also prepared from fresh leaf samples of the genome representations of tree ZG14. These genome representations were labeled and hybridised to new slides. Signal intensity values of the replicate hybridisations were plotted against each other in Microsoft Excel.

RESULTS

DNA Microarray Analysis

To consider the potential use of microarrays for fingerprinting *Eucalyptus* clones, a prototype DNA microarray chip was constructed with selectively amplified restriction fragments of pooled genomic DNA of an *E. grandis* full-sib family. The technique used to generate a genome representation of the full-sib family and of

each *Eucalyptus* tree employs the principle of AFLP (Vos et al. 1995). The complexity of each genomic DNA sample was reduced 16-fold by using +1/+1 selective nucleotides for PCR amplification of genomic restriction fragments. PCR amplicons prepared in this way ranged from 0.2 to 1.5 kb with an average insert size of 700 bp. Sequencing of 40 of the cloned PCR products revealed that there was a low proportion (17%) of "repeat" clones (i.e. clones with microsatellite or other simple repeat sequences, or multiple copies of the same genomic DNA fragment) in the *Eucalyptus* library generated (data not shown).

Proportion of polymorphic fragments useful for fingerprinting

To determine the proportion of polymorphic DNA fragments on the fingerprinting chip, tree ZG14 and 15 full-sib progeny were used in single dye experiments (Figure 2). While many of the array features were common (momomorphic) to all individuals (58%), or showed no hybridisation signal (15%), many (104 or 27%) were clearly polymorphic among individuals. However, only 55 of these spots (15%) were selected for further analyses. The analysis was limited to these 55 spots because clear threshold values (difference of 0.5 in relative intensity between two intensity classes) could be assigned for them (Figure 3A) and they were easily convertible into a binary scoring table (results not shown). In contrast, non-polymorphic spots, including both clearly monomorphic loci and loci that were not possible to score as either monomorphic or polymorphic (Figure 3B), exhibited a greater proportion of high relative intensity values. This can be attributed to the fact that monomorphic loci share the same signal intensities. Polymorphic spots for

which no clear threshold values could be assigned are responsible for the lower relative intensity values.

The CLUSTER software programme allowed us to visualise the relationships of the hybridisation profiles using TreeView (Eisen et al. 1998, Figure 4). The branching orders of duplicate experiments were all identical and duplicate experiments clustered as nearest neighbors. However, depending on which similarity metric setting was used, the overall branching order varied substantially. Since the Spearman correlation analysis provides a more conservative and reliable estimation of the relationship between hybridisation profiles (Murray et al. 2001), this correlation was used for data analysis. The dendrogram generated merely provides a means to visualize the relationship of fingerprints and should not be seen as representative of genetic relationships between the full-sib progeny.

All of the hybridisation profiles were unique and allowed unambiguous discrimination of the full-sib individuals. The probability of obtaining a particular 55- locus fingerprint is 2.7×10^{-17}, assuming no linkage among polymorphic spots. This provides an upper estimate of the discriminating power of our data. Randomly selected small subsets of polymorphic DNA fragments were used to determine that as few as 7 polymorphisms were sufficient to discriminate among full-sib progeny.

Reproducibility

To assess the reproducibility of the experimental procedure, replicate experiments were performed for nine individuals (Figure 2 and Figure 4). Signal intensities of the experimental replicates exhibited regression coefficients (R^2) ranging from 0.90

to 0.93 (Table 1). These are considered to reflect acceptable levels of reproducibility for microarray analysis (Hertzberg et al. 2001). These values were compared to the repeatability of binary scores obtained from the same hybridisations. Binary scores of replicate experiments were on average 1.5% higher than regression coefficients.

The regression of the hybridisation (normalized signal intensity) data obtained from two different sources of DNA (Figure 5) for the parent ZG14 revealed a linear regression coefficient (R^2) of 0.91. This was not significantly different from the regression coefficient obtained for the experimental replicates of the same tree ($R^2 > 0.93$), suggesting that independent DNA sampling did not introduce much additional experimental variance.

Validation of DNA polymorphisms

Two polymorphic DNA fragments (no 227 and 229) were analyzed by Southern hybridisation. When probe 227 of the genomic library was hybridised to a blot of the representations, trees 44D, 32A, 67D, TAG-S, 36E and 30B produced a band of 300bp in size, while a band of 430bp was detected for the other genotypes. The genomic Southern blot of probe 229 resulted in a band of 500bp in the case of trees 28C, ZG14, 60D, 17C, 10D, and a band of 300bp in size for the other genotypes. These RFLP banding patterns were converted to absence/presence of a band. These RFLPs were consistent with the bimodal hybridisation pattern observed for these two probes in the microarray experiment (Table 2).

Stripping and re-use of slides

Coefficients of determination, which are a measure of the correlation between two variables (experiments), were observed to be higher than 0.90 in replicate hybridisation experiments on stripped slides (data included in Table 1). This confirmed that re-used slides resulted in reproducible data. Although the signal intensities decreased on average by 10% after each successive hybridisation (Figure 2), spot signal intensities remained detectable and were quantifiable.

DISCUSSION

In this study we have shown that microarray technology can be used for genome-wide fingerprinting of closely related *Eucalyptus* trees. Several features of the DNA microarray technology make it attractive for this purpose. The DNA for hybridisation is prepared by selective PCR amplification of short restriction fragments. This means that <250 ng of total genomic DNA provides essentially unlimited starting material for future genotyping of the same trees. This technique, like AFLP analysis, also allows genomic fingerprinting of organisms such as *Eucalyptus* tree species with no prior DNA sequence information (Jaccoud et al. 2001). Most importantly, analysis of the polymorphic fragments is not restricted by the need for gel electrophoresis, and thousands of polymorphic loci in each tree genome, can potentially be analysed in a single assay. Gel electrophoresis in

contrast is limited in throughput and suffers from difficulties in precisely matching allelic variants of the same size on different gels (Ticknor et al. 2001).

Despite the recent progress that has been made towards the application of microarray technology for DNA fingerprinting and high-throughput genotyping in plants (Jaccoud et al. 2001, Borevitz et al. 2003), cross-hybridisation remains a problem. The highly repetitive DNA content of plant genomes undoubtedly results in cross-hybridisation of DNA fragments to printed probe DNA. This increases the overall spot intensity of many probes, and it masks potential polymorphisms. It has been demonstrated that small regions of similarity can lead to cross-hybridisation on oligonucleotide microarrays. Kane et al. (2000) found that in 50-mer oligonucleotide arrays, cross-hybridisation occurred between fragments of relatively low sequence similarity. This has also been observed on microarrays with PCR-based probes (Wren et al. 2002). In general, cross-hybridisation of many different genomic fragments will result in the conversion of polymorphic probes into monomorphic probes. However, a much more serious problem is presented by background segregation of a small number of strongly cross-hybridising fragments, which will result in mixed hybridisation patterns and incorrect marker phenotypes. This problem can be detected at the locus level in segregating progeny, but not in population or fingerprinting studies.

The prototype microarray chip developed in this study for fingerprinting *Eucalyptus* clones allowed for the discrimination among full sib progeny and thus

very closely related individuals. The hybridisation profiles obtained for *Eucalyptus grandis* individuals were highly repeatable ($R^2 > 0.9$), and allowed us to identify distinct intensity classes (bimodal intensity distributions) for 55 (14.3%) of the 384 printed probes (Figure 3). An additional 49 of the probes showed bimodal intensity distributions, but the overlap between the two intensity classes for these probes was inordinately great to easily assign them to presence or absence classes. The total proportion of bimodal probes (27%) and polymorphisms (14.3%) that could be scored was somewhat lower than the rate of polymorphisms often reported for gel-based AFLP markers in outcrossed *Eucalyptus* pedigrees (up to 50%, Myburg et al. 2003). The lower rate of scorable polymorphisms is most probably the result of cross-hybridisation obscuring polymorphic features. This is in addition to the "normal" inaccuracies introduced during labeling and hybridisation.

In an outcrossed pedigree, the majority of restriction fragment polymorphisms would be expected to segregate in testcross configuration (Aa:aa = 1:1), while a smaller proportion are expected to segregate in intercross configuration (AA:Aa:aa = 1:2:1 or 3:1). The majority of fragments will segregate as testcross fragments since a higher heterozygosity is expected in an outcrossing pedigree. Our pedigree set (15 full-sibs) was not sufficiently large to reliably distinguish between intercross and testcross segregation patterns, or to determine whether these fragments can be scored in a dosage dependent (co-dominant) fashion on microarrays. Therefore, the bimodal intensity distribution shown in Figure 3A probably contains a mixture of testcross and intercross fragments, which may explain the wide and high.

Signal intensity differences among genotypes can be compared across arrays using either single-dye or two-dye colour detection. The Diversity Array technique as described by Jaccoud et al. (2001) represents a two-dye approach. Differences among genotypes (presence or absence of fragments) are detected by comparing the Cy3 signal of each array element to the Cy5 signal of a reference (another genome representation, or a labeled vector fragment). Polymorphic spots show a bimodal distribution of log ratios relative to the reference. The use of a vector-based reference therefore provides an internal standard for each spot and a way to control for differences in the amount of DNA spotted on each array. However, if the same amount of DNA is spotted in each position across arrays, as can be expected for spots printed with the same pin, the value of the reference channel has to be balanced against the additional cost of labeling. Significant variation in printing across arrays was not observed, and therefore used a single-dye approach and normalized signal intensities rather than signal ratios. The normalised intensities were used to identify polymorphic spots based on their bimodal frequency distribution across individuals.

Reproducibility is essential in genotyping and fingerprinting. We tested for reproducibility of fingerprinting profiles at the experimental and biological level and found that the R^2 of normalized mean signal intensities was always higher than 0.90 in duplicate experiments, even when different sources of genomic DNA were used). The observed variability in signal intensities of 6 – 9 % between replicates of the same individual (in different labeling and hybridisation reactions) can be

ascribed to variability in the experimental process. Spot variability probably resulted from inaccuracies introduced in labeling, array hybridisations, signal detection and quantification, or low hybridisation signal. A higher frequency of errors at lower signal intensities was also observed due to signals being close to the background noise (Hertzberg et al. 2001). In comparison to the mean signal intensities, the (dominant) binary scores obtained from the same hybridisation data were more repeatable (>95%). This was due to the fact that correct scores could still be obtained when signal intensities varied within signal intensity classes, and due to the low occurrence of spots that varied sufficiently to be erroneously classified. In addition, the repeatability of the hybridisation profiles based on the 55 scored polymorphic probes was on average approximately 1.5% higher than that based on the full set of 384 probes (data not shown).

The power of microarray-based fingerprinting lies in its ability to compare different genomes at large numbers of loci, in a single assay. In this context, direct comparison of signal intensity profiles may allow accurate identification of individuals, if proper normalization procedures are followed. Our results suggest that binary scores based on underlying hybridisation patterns are only marginally more repeatable than the hybridisation data. However, binary (or ideally co-dominant) scores are required to determine allelic frequencies in target populations and in order to calculate probabilities of misidentification for forensic purposes. Binary scores are also required for linkage analysis in mapping pedigrees. Although the use of this technology for linkage mapping still remains to be tested,

our results suggest that the technology is useful for rapid genome-wide comparison of closely related germplasm. This study showed that the branching orders of replicate hybridisation fingerprints were all identical and replicate fingerprints all clustered as nearest neighbors. This allowed for the unambiguous identification of *Eucalyptus grandis* individuals and the identification of two unknown samples (included as blind test samples).

Microarray-based fingerprints may allow the identification of genomic regions shared between related individuals, or identification of genomic regions inherited from specific parents in outcrossed pedigrees. Borevitz et al. (2003) recently demonstrated the use of oligonucleotide probes to demarcate recombination events along chromosomes of recombinant inbred lines of *Arabidopsis*. In our case, map information is not available, but in the future the internal sequences of probes will be useful to link polymorphisms to a genome sequence when that becomes available for *Eucalyptus*. The clustering of probes into columns according to levels of similarity based on their hybridisation (or segregation) patterns across individuals suggests the presence of major linkage groups. This approach may allow ordering of polymorphic markers if the population size is increased adequately.

No dedicated software products are currently available to define hybridisation-based DNA fingerprints or to extract binary scores from hybridisation data. The majority of available microarray software is designed for two-color expression profiling studies. For single-color fingerprinting applications, such as the one used

in this study, the presence or absence of fragments (dominant scoring) or signal intensity (for co-dominant scoring) has to be determined to construct a fingerprint, and quality values need to be assigned to each data point to evaluate the reliability of the combined fingerprint. Kingsley et al. (2002) used the APEX (automated peak extraction) algorithm to measure spot intensities, and to determine whether a spot is "on" or "off". This algorithm has advantages over the software used in the present study, and should be considered for future work.

The long-term objective of the research presented in this study is to develop a larger array, or set of arrays, with informative probes that can be used for genome-wide fingerprinting of most commercially planted *Eucalyptus* tree species. This will require multiple rounds of selection of polymorphic probes within *E. grandis*, and selection of polymorphic probes in other species or interspecific mapping pedigrees. Such an array of polymorphic probes will be useful to saturate existing genetic linkage maps, and may also allow comparative mapping of many eucalypt genomes. Future fingerprinting arrays may be based on oligonucleotides residing in genes (Borevitz et al. 2003) or on genomic restriction fragments such as those cloned in this study.

REFERENCES

Botstein D, White R, Skolnick M, Davis R (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32: 314 – 331.

Borevitz JO, Liang D, Plouffe D, Chang H-S, Zhu T, Weigel D, Berry CC, Winzeler E, Chory J (2003) Large-scale identification of single-feature polymorphisms in complex genomes. Genome Res 13: 513-523.

Dolan PL, Wu Y, Ista LK, Metzenberg RL, Nelson MA, Lopez GP (2001) Robust and efficient synthetic method for forming DNA microarrays. Nucleic Acids Research 29(21): e107.

Eisen MB, Spellmann PT, Brown, PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci, USA 95: 14863 – 14868.

Eldridge KJ, Davidson J, Harwood C, van Wyk G (1993) Eucalypt Domestication and Breeding. Oxford Univ. Press, Oxford.

Epplen JT, Ammer H, Epplen C, Kammerbauer C, Mitreiter R, Roewer L, Schwaiger W, Steinle V, Zischler H, Albert E, Andreas A, Beuermann B, Meyer W, Buitkamp J, Nanda I, Schmid M, Nuernberg P, Pena SDJ, Poeche H, Sprecher W, Schartl M, Weising K, Yassouridis A (1991). Oligonucleotide fingerprinting using single repeat motifs: a convenient, ubiquitously applicable method to detect hypervariablity for multiple purposes. In: Burke T, Dolf G, Jeffreys AJ, Wolff R (eds) DNA fingerprinting approaches and applications, Birkhaeuser Verlag, Basel, pp 50 – 69.

Hertzberg M, Sievertzon M, Aspeborg H, Nilsson P, Sandberg G, Lundeberg J (2001) cDNA microarray analysis of small plant tissue samples using a cDNA tag target amplification protocol. The Plant Journal 25(4): 1 – 9.

Jaccoud D, Peng K, Feinstein D, Kilian A (2001) Diversity Arrays: a solid state technology for sequence independent genotyping. Nucleic Acids Research 29(4) : e25.

Kane MD et al. (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. Nucleic Acids Research 28: 4552 – 4557.

Keil M, Griffin AR (1994) Use of random amplified polymorphic DNA (RAPD) markers in the discrimination and verification of genotypes in Eucalyptus. Theor Appl Genet 89: 442 – 450.

Kingsley MT, Straub TM, Call DR, Daly DS, Wunschel SC, Chandler DP (2002) Fingerprinting closely related Xanthomonas pathovars with random nonamer oligonucleotide microarrays. Applied and Environmental Microbiology 68(12): 6361-6370.

Murray MG, Thompson WF (1980) Rapid isolation of high molecular weight plant DNA. Nucleic Acids Research 8: 4321 – 4325.

Murray AE, Lies D, Li G, Nealson K, Zhou J, Tiedje JM (2001) DNA/DNA hybridization to microarrays reveals gene-specific differences between closely related microbial genomes. Proc Natl Acad Sci 98 (17): 9853-9858.

Myburg AA, Griffin AR, Sederoff RR, Whetten RW (2003) Comparative genetic linkage maps of *E. grandis, E. globulus* and their F_1 hybrid based on a double pseudo-backcross mapping approach. Theor Appl Genet (in press).

Nybom H (1991) Applications of DNA fingerprinting in plant breeding. In: Burke T, Dolf G, Jeffreys AJ, Wolff R (eds) DNA fingrprinting approaches and applications, Birkhaeuser Verlag, Basel, pp. 294 – 331.

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270: 467 – 470.

Smith S, Beavis W (1996) Molecular marker assisted breeding in a company environment. In: The Impact of Plant Molecular Genetics, ed. B.W.S. Sobral, chap15. Birkhauser, Boston.

Ticknor LO, Kolsto A-B, Hill KK, Keim P, Laker MT, Tonks M, Jackson PJ (2001) Fluorescent amplified fragment length polymorphism analysis of Norwegian *Bacillus cereus* and *Bacillus thuringiensis* soil isolates. Appl Environ Microbiol 67: 4863 – 4873.

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Fritjers A, Pot J, Peleman J, Kuiper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Research 23: 4407 – 4414.

Weber J, May P (1989) Abundant class of human DNA polymorphisms which can be typed using the polymearse chain reaction. Am J Hum Genet 44: 388 – 396.

Weising K, Ramser J, Kaemmer D, Kahl G, Epplen JT (1991) Oligonucleotide fingerprinting in plant and fungi. In: Burke T, Dolf G, Jeffreys AJ, Wolff R (eds) DNA fingrprinting approaches and applications, Birkhaeuser Verlag, Basel, pp 312 – 329.

Williams J, Kubelik A, Livak K, Rafalski J, Tingey S (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Research 18: 6531 – 6535.

Wren JD, Kulkarni A, Joslin J, Butow RA, Garner HR (2002) Cross-Hybridization on PCR-Spotted Microarrays. IEEE Engineering in Medicine and Biology. March/ April 71 – 75.

Table 1. Repeatability of hybridization profiles and binary scores. R^2 values are based on two separate labeling reactions and hybridisations starting from a single genome representation of each individual.

Eucalyptus individual no	R^2 [a] Hybridization profile	Repeatability of binary scores [b]
ZG14 (parent tree)	93.53	98.18
ZG14 (parent tree – biological replicate)[c]	91.47	94.55
TAG-5 (relative)	91.72	96.37
18C	92.34	96.37
28C	91.79	94.55
53B	93.95	98.18
36E	92.52	96.37
30B	93.09	96.37
67D	91.86	94.55
74C	93.97	98.18

[a] Based on the spot intensities in two replicate experiments of all 384 features on the array.

[b] Proportion of polymorphic probes (55 total) with same binary score across experimental replicates [1-(number of misscores/55)]x 100%

[c] For tree ZG14, in addition to a direct experimental replicate, an independently obtained DNA sample and genome representation was used as biological replicate.

Table 2. Hybridization patterns of individual RFLP alleles and microarray features. Hybridization patterns were only determined for replicated individuals (see Table 1).

		ZG14	74C	18C	28C	53B	36E	30B	67D	TAG-5
Probe 227	RFLP allele **(300 bp)**	_	_	_	_	+	+	_	+	+
	Microarray hybridisations	_	_	_	_	+	+	_	+	+
Probe 229	RFLP allele **(350 bp)**	_	_	+	+	+	+	+	+	+
	Microarray hybridisations	_	_	+	+	+	+	+	+	+

Figure 1. Schematic representation of the microarray-based genotyping method used in this study. Note that a pooled DNA sample was used to prepare the genome representation that was printed on the array. Also, the length of the selectively amplified restriction fragments determine the number of incorporated dye molecules per fragment, and therefore the average intensity of the corresponding spot on the array.

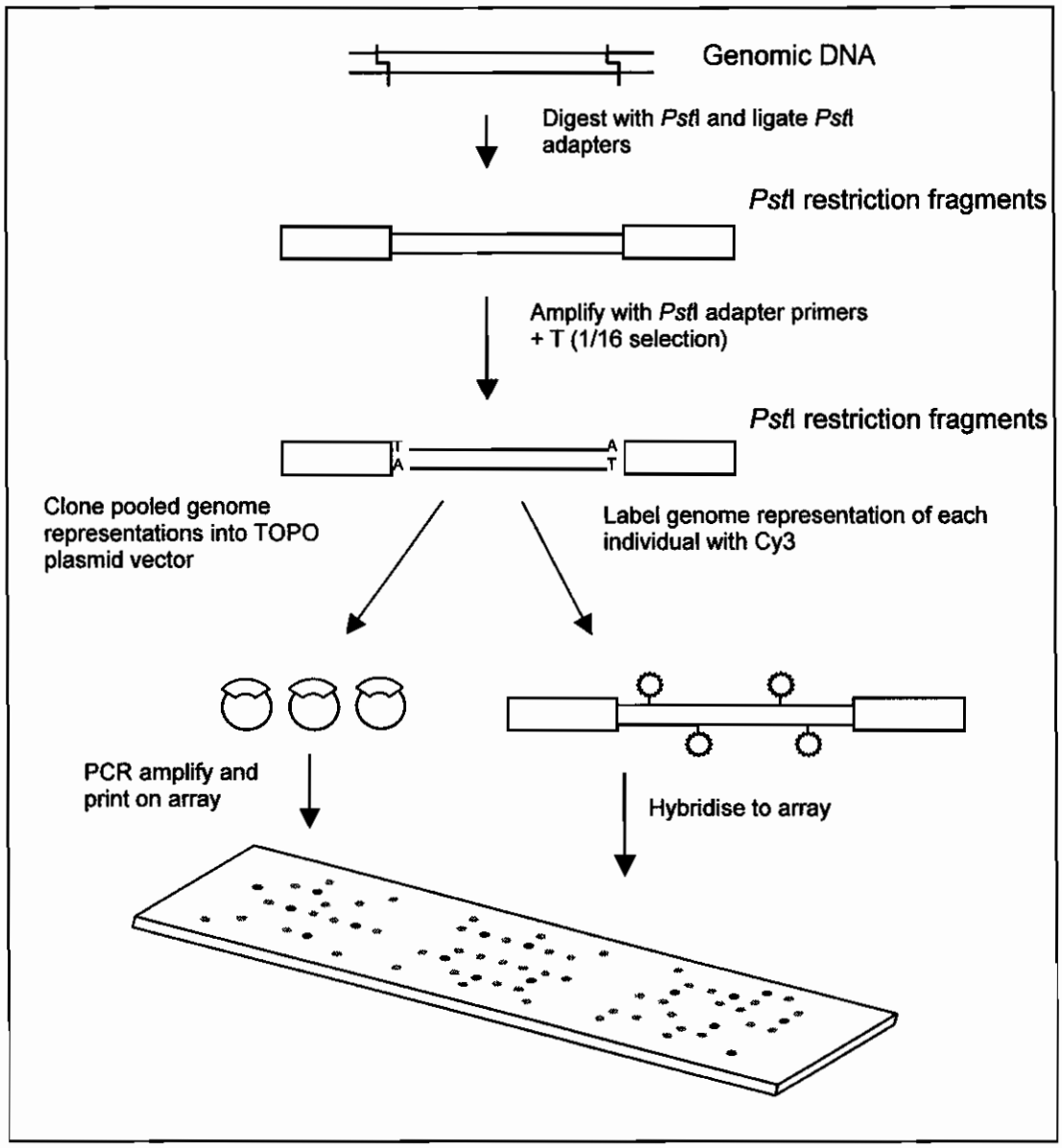

Genomic DNA
Digest with *Pst*I and ligate *Pst*I adapters
*Pst*I restriction fragments
Amplify with *Pst*I adapter primers + T (1/16 selection)
*Pst*I restriction fragments
Clone pooled genome representations into TOPO plasmid vector
Label genome representation of each individual with Cy3
PCR amplify and print on array
Hybridise to array

Figure 2. Microarray hybridization patterns of two different *Eucalyptus* individuals on the same section of the slide. Each column represents four replicates of the same spot. (A) Hybridization fingerprint of *Eucalyptus* individual 67D and (B) parent ZG14. C and D are hybridization of the same individuals on replicate, stripped arrays.

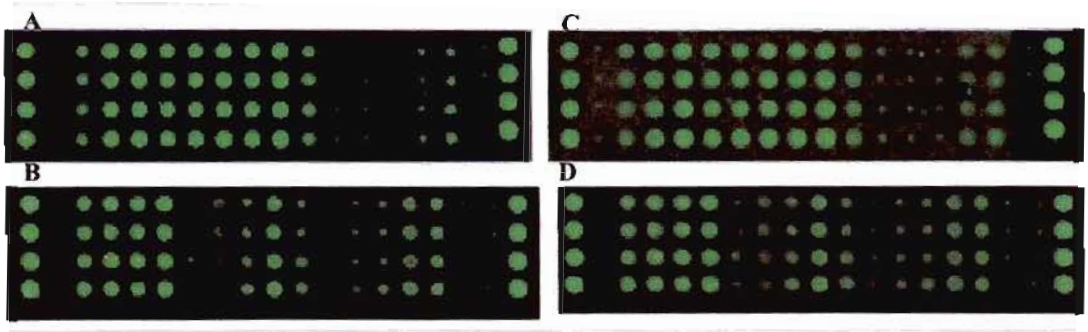
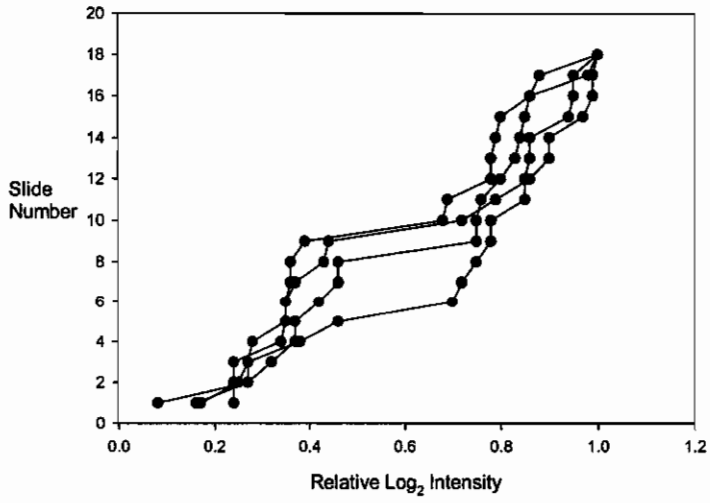


Figure 3. Examples of signal intensity distributions of log-transformed hybridisation data among 17 *Eucalyptus* individuals. (A) Distribution of relative (normalized) log intensities of four random polymorphic fragments that show a clear bimodal distribution across slides. (B) Non-polymorphic spots show a unimodal distribution.

A


Slide Number
0
2
4
6
8
10
12
14
16
18
20
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Relative Log2 Intensity

B

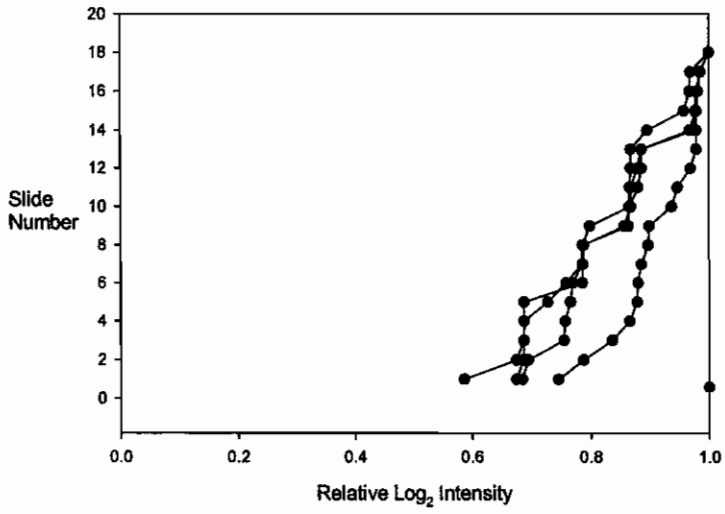

Slide Number
0
2
4
6
8
10
12
14
16
18
20
0.0
0.2
0.4
0.6
0.8
1.0
Relative Log2 Intensity

Figure 4. TreeView (Eisen et al. 1998) representation of relationships of hybridization profiles among 17 *Eucalyptus* individuals based on microarray analysis with the 384-probe array. Columns represent hybridization profiles of individuals (or replicates) and rows represent the mean log intensities for labeled DNA/DNA hybridisations across individuals. *Red* and *green* bars indicate high and low mean log intensity values, *black* bars indicate intermediate values and *grey* bars show missing data. Nine of the hybridisations were performed in replicate (indicated as REP). The replicate for ZG14 is a biological replicate, i.e. starting from independently obtained leaf samples of the same tree.

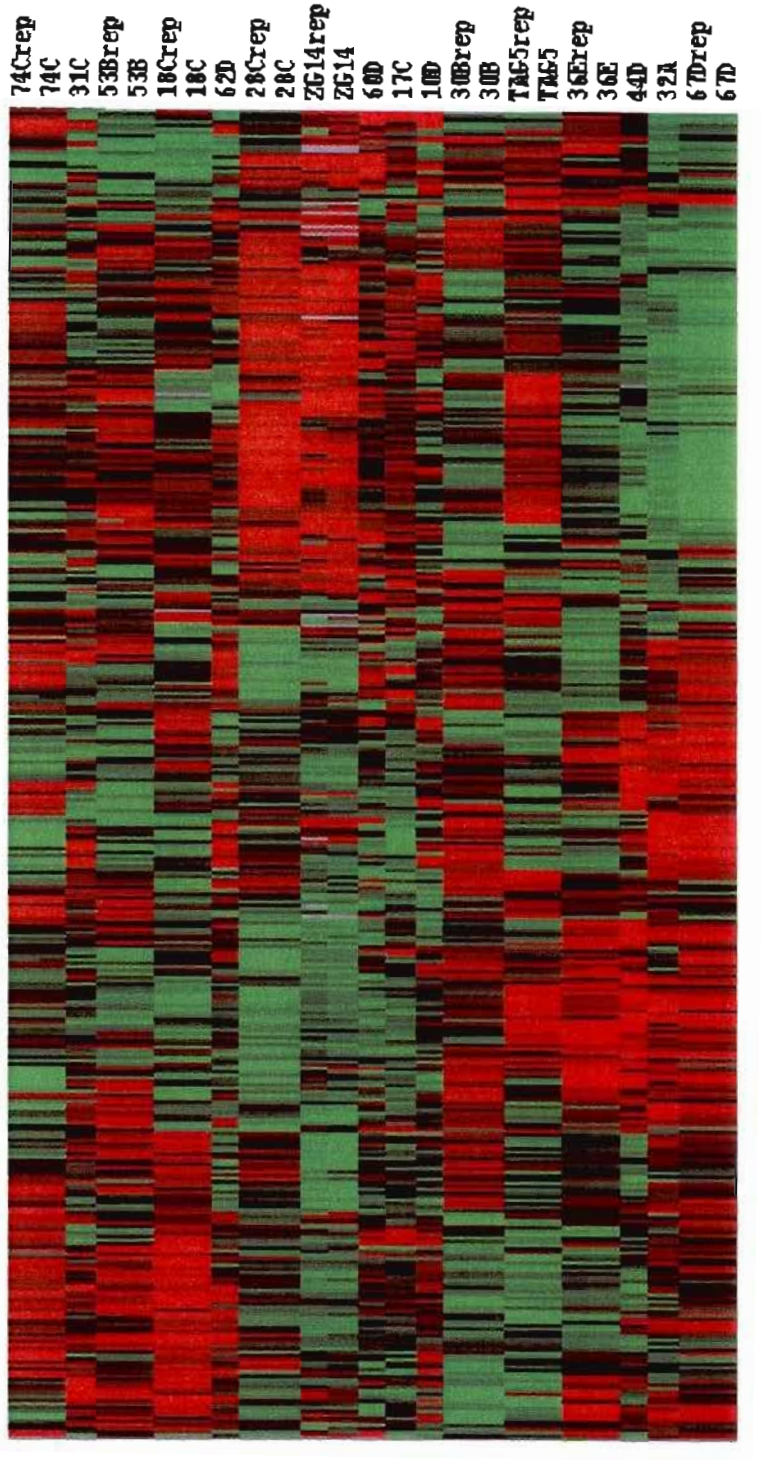
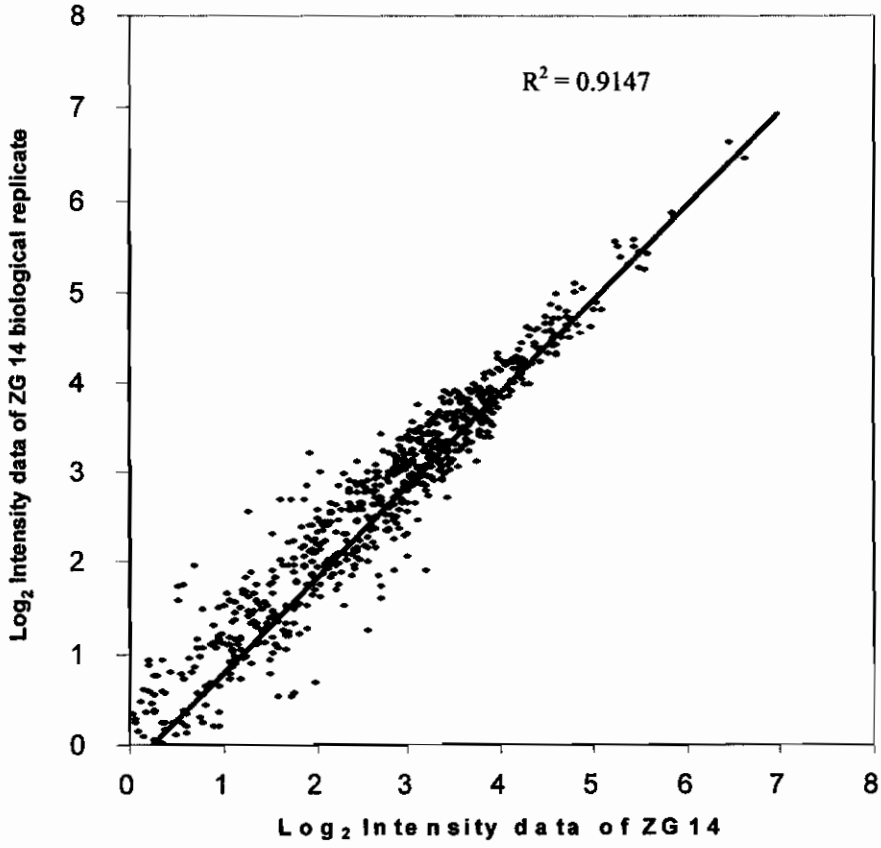

74Crep
74C
31C
53Brep
53B
18Crep
18C
62D
28Crep
28C
ZG14rep
ZG14
60D
17C
10D
30Brep
30B
TM5rep
TM5
36Erep
36E
44D
32A
67Drep
67D

Figure 5. Log plot of the microarray hybridisation signals of *Eucalyptus* individual (ZG14). The signal intensity obtained with ZG14 (*x*axis) was plotted against its biological replicate.


$R^2 = 0.9147$
Log_2 Intensity data of ZG 14 biological replicate
Log_2 Intensity data of ZG 14
0
1
2
3
4
5
6
7
8